IMPLEMENTATION OF SUPPORT VECTOR MACHINE ALGORITHM FOR THE CLINICAL DIAGNOSIS OF BREAST CANCER

Ugwu Edith Angela
Department of Computer Science, Enugu State University of Science and Technology, (ESUT), Agbani, Enugu State Nigeria

Article Info
Received: 24/3/ 2024
Revised: 17/4/2024
Accepted 4/5/2024

Corresponding Author's Email: edith.ugwu@esut.edu.ng

Corresponding Author's Tel:
+234 803 784 2144

ABSTRACT
Over the years, the increasing incidence of mortality among women due to breast cancer has posed a significant challenge, necessitating urgent solutions. While numerous studies have attempted to address this issue, the lack of a clear definition of success limits the reliability of existing solutions. This paper focuses on implementing the support vector machine algorithm for the clinical diagnosis of breast cancer. A quantitative research design approach was employed, involving methods such as data collection, imputation, and transformation. The model was developed using the Python programming language and subsequently evaluated using well-defined success criteria, including accuracy, False Detection Rate (FDR), True Positive Rate (TPR), and Receiver Operating Characteristic (ROC) curve analysis. The model achieved an accuracy of 91.89% and minimized the FDR to 8.11%, indicating its ability to reduce false positives and enhance reliability. Additionally, it attained a ROC value of 0.65 and a perfect TPR of 1.00, demonstrating its effectiveness in identifying actual positive cases. Collectively, these metrics underscore the promising performance of the model and its potential to facilitate early breast cancer diagnosis. Validation of the model employed a comparative approach, which revealed its superior reliability compared to existing systems. The new model proposed in this research is recommended for developing software-based diagnostic systems for breast cancer detection and management. Further studies should focus on practical validation of the model through real-world experimentation.

Keywords: Breast Cancer; Support Vector Machine (SVM); Machine Learning; Diagnostic System

1. INTRODUCTION

All over the world, breast cancer stands among the major contributor of mortality rate for women, and has remained a major global concern. This problem has affected numerous lives socially and health-wise, resulted to loss of lives, and necessitates timely and accurate detection or effective management. Overall, breast cancer can be classified into two types which are the non-invasive and invasive respectively (Medjahed and Benyettou, 2013). In the non-invasive types, the cells are confined to the ducts and do not invade surrounding tissues of the breast, with the Ductal Carcinoma In-Shitu (DCIS), the most common of the problem, with average of 90% of recorded cases (Ershler, 2005). The Invasive Breast Cancer (IBC) on the other hand invades the surrounding tissues of the breast regions (Hallen et al., 2015).

According to Ene and Ekwote (2023), early detection of breast cancer is a vital to enhance the chances of effective diagnosis and control the rapid growth of the problem. Traditional methods of detecting breast cancer include mammography, clinical breast examination, self-examination, ultrasound, magnetic resonance imaging and biopsy (Shravya et al., 2019). While these methods have contributed greatly to the early detection of breast cancer, the issues of false negative have remained a major problem, as eventually in most of the cases, decision are made based on human judgment, which in some cases may be prone to error, inaccurate, or false alarm. Most recently the application of Artificial Intelligence (A.I) has been explored in medical fields as a clinical assistance mechanism to improve decision making. For instance, Keles and Kaya (2019) and Akbugday (2019) trained K-Nearest Neighbor (K-NN, Naïve Bayes (NB) and Support Vector Machine (SVM) to generated three models or the classification of breast cancer. Overall SVM was identified as the best with accuracy averaging 96.93%. In another paper, Sinthia et al.

(2015) trained Neural Network (NN) and Linear Regression (LR) for the classification of breast cancer and recommended NN as the best with 94.2% accuracy, while Chaurasia and Pal (2016) reported 97.13% accuracy with SVM, while Khan et al., (2018) reported an average of 74.44% after training and evaluating SVM and NB.

While these studies all contributed to early detection of breast cancer, SVM was identified as one of the most effective machine learning algorithms for classification of breast cancer; however, critical gap lies in the definition of the model success (Preda and Bellomi, 2017). This is because only accuracy was majorly considered as the evaluation of the models, however, other metrics like recall, precision, true positive and false positive when applied to evaluate these models, would provide more insights on their performance and trustworthiness towards early detection and classification of breast cancer. To this end, this study seeks to presents a more reliable model trained with data of early breast cancer features utilizing support vector as the machine learning algorithm. This will achieve will provide a more reliable classification model for the implementation of clinical decision system for diagnosis of breast cancer.

2. RESEARCH METHODOLOGY

The research methodology employed in this study involved a systematic approach to investigate breast cancer. The process began with the collection of relevant data pertaining to breast cancer. Subsequently, data processing was conducted, encompassing the identification and handling of missing values through a comprehensive search using Gretel toolbox. To enhance the quality and utility of the data, a crucial step involved the application of the frequency doubling transformation techniques.

Following the data preprocessing phase, the research proceeded to train a Support Vector Machine (SVM) using the prepared dataset. The SVM, a machine learning algorithm, underwent a training process to learn patterns and relationships within the data. This involved exposing the SVM to labelled examples from the dataset, allowing it to iteratively optimize its parameters. The objective of this training phase was to enable the SVM to create a predictive model capable of accurately classifying breast cancer instances based on the input features present in the dataset.

2.1 Research Design

The research design approach used is the quantitative approach and the reason was because the research deals with statistical data and also machine learning algorithms. The approach comprises three key objectives aimed at advancing the understanding of breast cancer and improving early detection methodologies. The first objective focuses on parameter acquisition by collecting data that encapsulates the early features of breast cancer. This involves a comprehensive exploration of relevant variables and characteristics that are indicative of the disease's initial stages. Through rigorous data collection, the research seeks to lay the foundation for a better understanding of the parameters that define the early detection of breast cancer.

The second objective involves the utilization of a Support Vector Machine (SVM) learning algorithm to harness the collected breast cancer data. The SVM is employed as a powerful tool for training a diagnostic system designed to detect breast cancer at its early stages. By using this machine learning techniques, the researcher aimed to develop a good model that can effectively identify patterns and relationships within the breast cancer data, ultimately leading to its early breast cancer detection ultimately.

The final phase of the research design centres on the evaluation and validation of the developed model. This critical step involves assessing the performance of the SVM-based diagnostic system and validating the results obtained. Rigorous evaluation metrics and validation procedures are employed to ensure the reliability and accuracy of the model considering parameters such as accuracy, precision confusion matrix, receiver operator characteristics curve.

2.2 Data collection

The breast cancer data used for this research was collected from Kaggle open-source repository. The data size contained 5000 features of breast cancer collected across diverse ages women between 30 to 70years. The data contains 16 main attributes of breast cancer and the choice of this attribute selection was through interaction with domain expert from the Enugu State University Teaching Hospital, Park-lane. The table 1 presents the data description. The breast cancer dataset attributes includes various attributes that provide information about the patients and the characteristics of their cancer. Numeric attributes such as "Age," "Survival Months," "Tumor Size," "Regional Node Examined," "Reginol Node Positive," and "Differentiate" offer quantitative insights into patient age, survival duration, tumor size, and examination results. Categorical attributes such as "T Stage," "Marital Status," "Race," "6th Stage," "N Stage," "Grade," and "A Stage" capture qualitative information related to the tumor stage, marital status, race, cancer stage, lymph node involvement, and tumor grade. Binary attributes like "Status," "Progesterone Status," and "Estrogen Status" represent binary outcomes or characteristics. These attributes collectively contribute to a

comprehensive understanding of the breast cancer dataset, encompassing both quantitative and qualitative factors relevant to the disease and patient profiles.

Table 1: Data description

Field	Data Description	Data Type
Age	Age of the patient at the time of diagnosis.	Numeric
Survival Months	Duration of survival measured in months.	Numeric
Tumor Size	Size of the tumor, typically measured in a unit such as centimeters.	Numeric
Regional Node Examined	Number of regional lymph nodes examined during diagnosis.	Numeric
T Stage	Categorization of the tumor stage.	Categorical
Reginol Node Positive	Number of regional lymph nodes testing positive for cancer.	Numeric
Marital Status	Current marital status of the patient.	Categorical
Status	Current health status of the patient (e.g., Alive or Deceased).	Binary
Race	Racial background or ethnicity of the patient.	Categorical
6th Stage	Cancer stage classification, possibly referring to the 6th edition.	Categorical
N Stage	Categorization of nodal stage in cancer.	Categorical
Grade	Categorization of tumor grade.	Categorical
Differentiate	Degree of tumor differentiation.	Categorical
A Stage	Additional stage-related information.	Binary
Progesterone Status	Progesterone receptor status of the tumor.	Binary
Estrogen Status	Estrogen receptor status of the tumor.	Binary

3. SUPPORT VECTOR MACHINE (SVM)

A Support Vector Machine (SVM) is a powerful machine learning algorithm commonly used for the classification of the breast cancer problem in this research. Its strength lies in its ability to find an optimal hyper-plane that separates data points into distinct classes, aiming to maximize the margin between these classes. When the SVM is trained, the decision boundary which is the hyper-plane is generated, which is now used to make decision such as classification. When a new data point is introduced, SVM evaluates which side of the hyperplane it falls on. The sign of the result determines the predicted class of the data point.In essence, SVM uses the position of data points in relation to the hyperplane to classify them into different classes.

3.1 Training of the SVM

To train the SVM, the data processed was imported into the model, while the hyper-parameters such as kernel, regularization parameter (C), width and bias function are optimally adjusted until the optimal hyper-plane is generated. Before the training process, the data are divided into test, train and validation sets. The test set which are the true values are used to test the trained model which are the predicted values, while the disparities between then is the error, which informed that optimization process of the hyper-parameters. This process continued iteratively until the breast cancer detection model is generated as depicted in the figure 1.

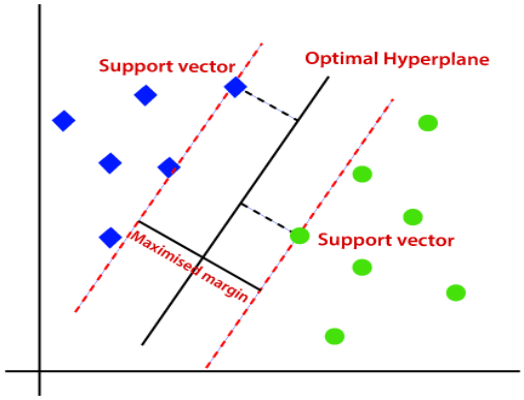

Figure 1: The trained SVM model

The figure 1 presented the trained SVM wit hyper-plane which is used in making decision on the detection of breast cancer. When new data is entered to the model, it uses the hyper-plane to decide where which class the data belongs to in the vectors and then predict the health status of the patient.

4. THE BREAST CANCER DETECTION MODEL

The breast cancer detection model generated with Support Vector Machines (SVM) is a powerful tool achieved in this research for the classification of breast cancer diseases, to facilitate medical diagnostics. This SVM based model proves

particularly effective in distinguishing between variant features of breast cancer, by creating an optimal hyper-plane that maximizes the margin between the two classes which are normally positive and negative (implying breast cancer and non-breast cancer). When test data is imported, the hyper-plane makes the boundary decision and then class the test data falls is the used to make the classification decision.

5. IMPLEMENTATION

The model was implemented using Python programming language facilitated by A.I Programming framework. First, the scikit-learn library, a powerful tool for machine learning tasks was utilized to initialize the data. The breast cancer dataset from scikit-learn was employed, consisting of features extracted from medical images and corresponding labels indicating tumor malignancy. The dataset was then split into training and testing sets using the **train_test_split** function. To enhance the model's performance, the features were standardized using the **StandardScaler** to ensure all features were on the same scale. Subsequently, a Support Vector Machines (SVM) classifier, accessible through the **SVC** class in scikit-learn, was employed to train on the standardized training data. Finally, the model's accuracy was evaluated on the testing set using the **accuracy_score** metric from scikit-learn, providing a measure of its performance in accurately classifying breast tumors.

5.1 Performance evaluation

The performance evaluation of the breast cancer detection model, generated considered key metrics such as Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) measure the trade-off between sensitivity and specificity, illustrating the model's discrimination ability. Accuracy (ACC) provides a general assessment of correctness, while the confusion matrix breaks down predictions into true positives, true negatives, false positives, and false negatives, offering a detailed overview of the model's performance. Precision evaluates the accuracy of positive predictions, crucial in minimizing false positives. A holistic understanding of these metrics is essential to assess the SVM model's efficacy in early breast cancer detection, ensuring a balance between sensitivity, specificity, and overall accuracy.

5.2 System Block Diagram

The system block diagram in figure 2 composed of the data collection of breast cancer information and then processed using the missing value handing through the mean imputation approach and the transformation approach using the frequency double transformation technique. The aim was to ensure compatibility of the model with the features vectors. The transformed data was used to train an SVM algorithm, through the optimization of the hyper-parameters such as kernel, regularization and weight, to generate the decision boundary applied for the breast cancer detection.

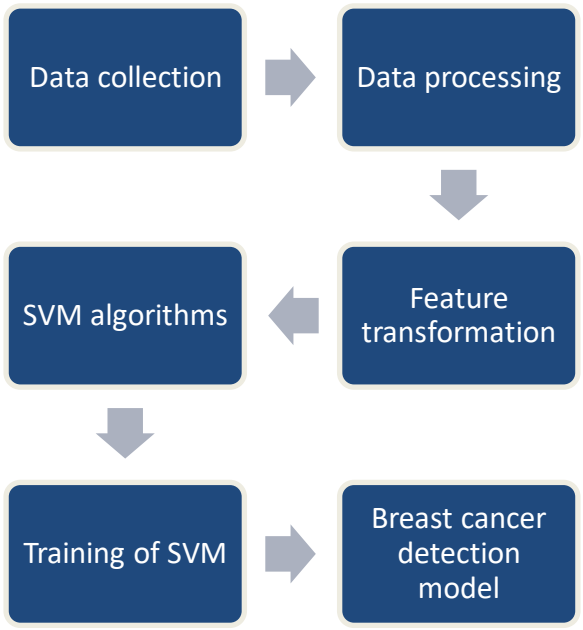

Figure 2: Block diagram of the breast cancer detection model

6. RESULTS AND DISCUSSION

This section begins with the result of the data visualization process. This was achieved using distribution of histogram to plot the relationships between certain key attributes of the breast cancer data, showing the frequency of the breast cancer occurrence among women of diverse ages of 30-70 as depicted in the figure 3;

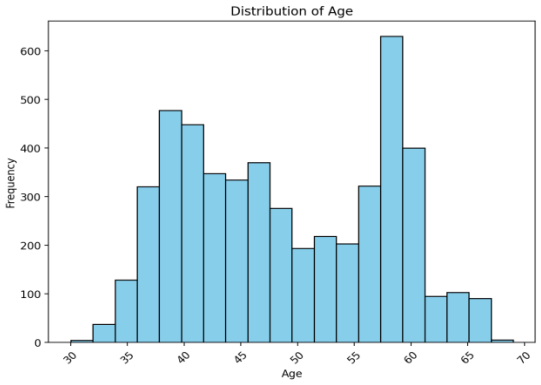

Figure 3: Distribution of the data features

The figure 3 applied histogram to shows that distribution of women who suffers breast cancer

across their ages. From the result it was observed that the age range of women who suffers this breast cancer the most is 58years. What this means is that this model developed will be recommended specifically for women mostly within this age bracket, to check and detect if there are signs of breast cancer and then manage early. Another main age of women who suffer this problem is 38years. A woman within this age range also suffers breast cancer issues a lot, and presents the need for clinical diagnosis and medical checkup frequently within this age. This may also suggest that women after child birth may have risk of breast cancer and information collected from domain expert showed that at 38years, most women are done with child birth. In the next result, the frequency double transformation approach applied for the visualization of the tumor size against the frequency of occurrence.

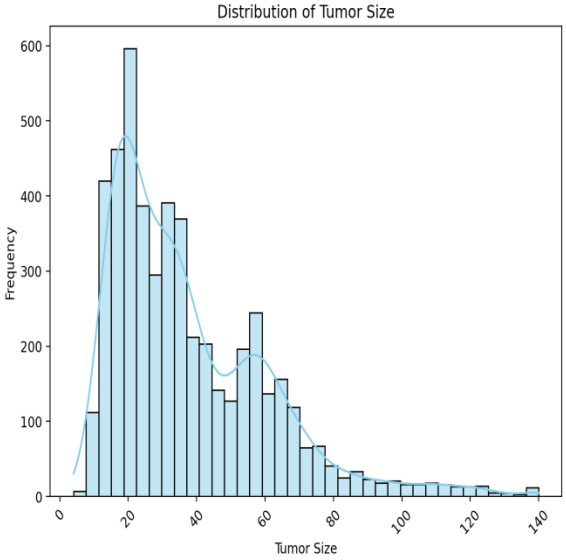

Figure 4: Distribution of tumor size

The figure 4 presents the distribution of tumor sizes across the diverse population of women with breast cancer. The result showed that breast cancer among mostly have tumor size of 20 or approximately 20. To this end, early detection is necessary to facilitate the management o this cancer and control the growth before it leads to a more critical health issue. The result of the data transformation process was reported in the figure 6;

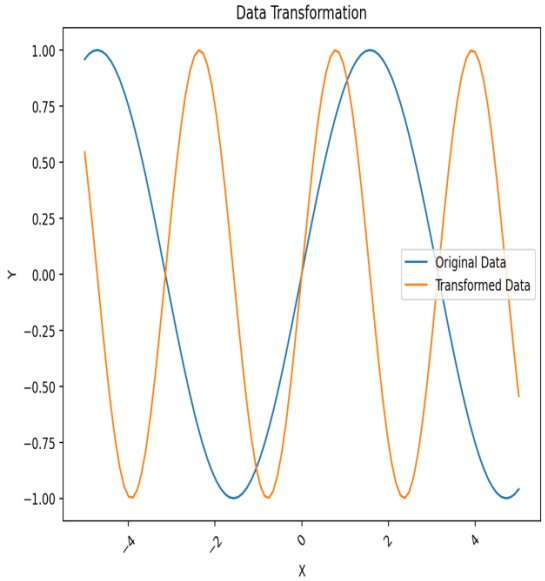

Figure 5: Frequency double transformation result

The figure 5 depicted the result of the frequency transformation approach used for the dimensionality reduction of the processed data before training the SVM model. This transformation method manipulates the frequency content of the cancer through the duplication of the frequency components within the data in a range of -4 and 4 (what this mean is that the data values will be compressed within the range of -4 and 4), while maintaining the data quality. Having transformed the data, it was loaded to SVM for training and the results presented using the confusion matrix in figure 6;

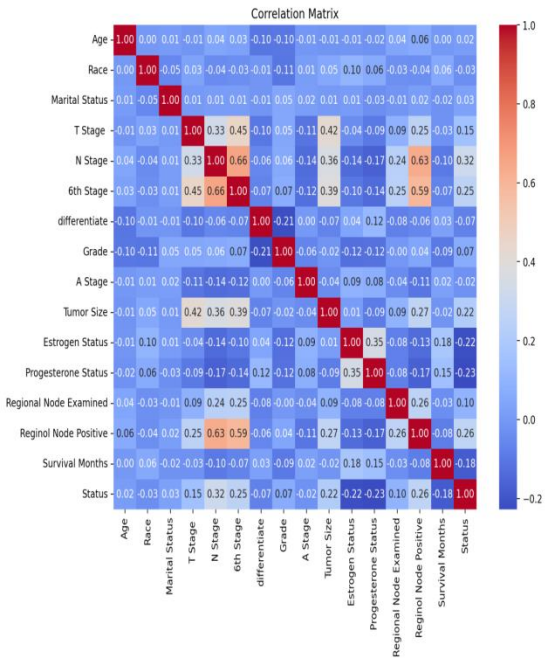

Figure 6: Confusion matrix result of the training process

The figure 6 presented the confusion matrix used for the evaluation of the model training. During this process, the feature of the breast cancer attributes are used to test the predicted value which is the trained model and then result measured between the range of 0 to 1, considering true positive (TP) and false positive (FT). From the confusion matrix, it was observed that overall for each attributes, the TP is 1.00 and FT is 0.00. What this mean is that the hyper-plane of the SVM model for the classification of breast cancer was able to correctly decide which class each of the data belongs correctly with high success rate. To measure the ROC performance, the relationship between the TP and FP was applied. The aim is to achieve a ROC value equal or approximately 1, which implied the efficiency rate of the model during applicability for the detection of breast cancer. The result was presented in the figure 7;

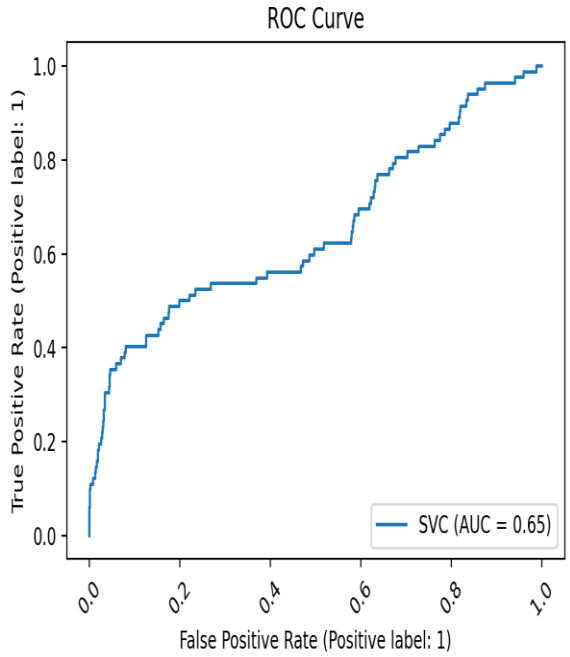

Figure 7: ROC result of the breast cancer detection model

The figure 7 showed that the area under the ROC curve reported 0.65 score. What this mean is that the ability of the model to correctly detect breast cancer and also detect when there is no breast cancer is 65% success. This result indicated that the model generated is good, a it was able to correctly detect the breast cancer problem with a high score, this indicating the effectiveness of SV for classification problem. Additionally, to measure the accuracy of the model, the figure 9 was applied.

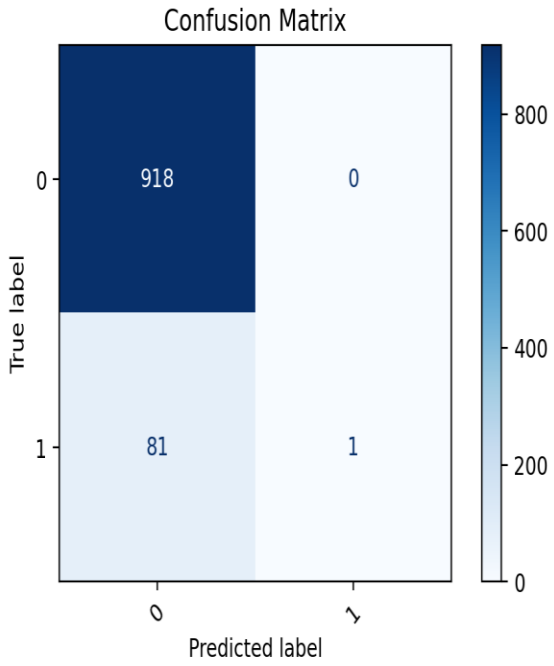

Figure 8: Accuracy of the breast cancer model

The figure 8 presented the accuracy score of the model using another confusion matrix differently. In this case the test data which is 1000 breast cancer samples was used to test the trained SVM model, and the result showed that the hyper-plane was able to correctly classify 918 as breast cancer and 81 as non-breast cancer; thus reporting a classification accuracy of 91.89% and false detection rate of 8.11%. This means that the model was able to successfully classify breast cancer correctly, thus collaborating with the earlier claim of the ability of SVM to correctly solve classification problem.

7. CONCLUSION

This study focused on the development of a diagnostic model for the detection of breast cancer using machine learning, specifically SVM algorithm. The SVM-based breast cancer detection model generated in this study presented a robust solution for medical diagnostics. The high accuracy, coupled with low false detection rates and strong ROC and true positive rate values, positions the model as a valuable tool in clinical settings. The emphasis on clear success definition is crucial in healthcare applications, and the model's ability to achieve this while maintaining a high true positive rate is noteworthy. This indicates the model's potential as an effective tool for early breast cancer diagnosis, showcasing its reliability in clinical applications.

8. REFERENCES

Akbugday B., (2019) Classification of Breast Cancer Data Using Machine Learning Algorithms. 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.

Chaurasia V.,& Pal S., (2016) Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. (FAMS 2016) 83 (2016) 1064 – 1069

Ene P., & Ekwote K., (2023)Development of clinical decision system for breast cancer diagnosis using machine learning technique. [J]AJSET; Vol 8(4), 2023

Ershler W., (2005) The Influence of Advanced Age on Cancer Occurrence and Growth. In: Balducci L., Extermann M, editors. Biological Basis of Geriatric Oncology. Springer US; 2005; 124: 75-87.

Hallen S., Hootsmans N., Blaisdell L., Gutheil C.,&Han P., (2015) Physicians perceptions of the value of prognosticmodels: the benefits and risks of prognostic confidence. HealExpect2015;18:2266e2277.https://doi.org/10.1111/hex.12196.

Keles M., & Kaya, (2019) Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study. Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

Khan S., Ullah N., Ahmed I., Ahmad I., & Mahsud M., (2018) MRI imaging, comparison of MRI with other modalities, noise in MRI images and machine learning techniques for noise removal: A review. Curr. Med. Imaging **2018**, 15, 243–254.

Medjahed T.,&BenyettouA., (2013) Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. International Journal of Computer Applications 62 (1), 2013

Preda, L.,&Bellomi M., (2017) On the Feasibility of Breast Cancer Imaging Systems at Millimeter-aves Frequencies. IEEE Trans. Microw. Theory Technol. **2017**, 65, 1795–1806.

Shravya K., &Shaik S., (2019) Prediction of Breast Cancer Using Supervised Machine Learning Techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, April 2019.

Sinthia P., Devi R., Gayathri R.,& Sivasankari R., (2017) Breast Cancer detection using PCPCET and ADEWNN", CIEEE' 17, p.63-65